Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 707580, 11 pages
doi:10.1155/2009/707580

*Research Article*

# Integrating Multiple Microarray Data for Cancer Pathway Analysis Using Bootstrapping K-S Test

**Bing Han,[1] Xue-Wen Chen,[1] Xinkun Wang,[2] and Elias K. Michaelis[2]**

[1] *Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA*
[2] *Higuchi Biosciences Center, University of Kansas, 2099 Constant Avenue, Lawrence, KS 66047, USA*

Correspondence should be addressed to Xue-Wen Chen, xwchen@ku.edu

Received 14 January 2009; Accepted 4 March 2009

Recommended by Dechang Chen

Previous applications of microarray technology for cancer research have mostly focused on identifying genes that are differentially expressed between a particular cancer and normal cells. In a biological system, genes perform different molecular functions and regulate various biological processes via interactions with other genes thus forming a variety of complex networks. Therefore, it is critical to understand the relationship (e.g., interactions) between genes across different types of cancer in order to gain insights into the molecular mechanisms of cancer. Here we propose an integrative method based on the bootstrapping Kolmogorov-Smirnov test and a large set of microarray data produced with various types of cancer to discover common molecular changes in cells from normal state to cancerous state. We evaluate our method using three key pathways related to cancer and demonstrate that it is capable of finding meaningful alterations in gene relations.



## 1. Introduction

Microarray technology, monitoring mRNA abundance of tens of thousands of genes simultaneously, provides an efficient tool to characterize a cell at the molecular level. It has been applied to a variety of research areas, ranging from biomarker detection [1, 2] to gene regulatory networks [3–5] and cancer classification [6–8]. When applied to cancer research, microarray technology typically measures gene expressions of cancer and normal tissues or different types of cancer. One important area in microarray-based cancer research is to identify genes that are differentially expressed between cancerous and normal cells and to discover diagnostic and prognostic signatures in order to predict therapeutic responses. Over the years, many statistical methods for the identification of differentially expressed genes have been developed, and most of them focused on the expression analysis of individual genes [9–15]. However, the simple list of individual differentially expressed genes can only tell us which genes are altered by biological differences between different cell types and/or states. It cannot explain the reasons for the significant alterations in gene expression levels and the effects of such changes on other genes' activities. It is well known that in a biological system genes interact with each other forming various biological pathways in order to carry out a multitude of biological processes. To better understand the roles of these differentially expressed genes and their interactions in a complex biological system, a comprehensive pathway analysis is needed. Since the identification of biological pathways is significantly influenced by those differentially expressed genes from different datasets or different statistical methods [16, 17], we reason here that an integration of multiple cancer microarray datasets and identification of the most common pathways from these data would reveal key relationships between crucial genes in carcinogenesis. Our focus on the interactions and pathways of cancer-related genes is important since changes in gene relations and key pathways are more relevant to carcinogenesis than individual genes alone.

Several statistical methods have been proposed for the analysis of differential gene coexpression patterns. Li [18] observed differences of gene coexpression patterns in different cellular states and attributed these changes in gene coexpression patterns to some third set of influential genes.

Lai et al. [19] proposed a similar method to identify differential gene-gene coexpression patterns in cells from normal state to cancerous state. However, these methods often perform the analyses on one single microarray dataset and typically generate unreliable results; the results from different microarray datasets and various statistical methods could hardly overlap using these methods [20, 21]. Therefore, the confidence level for discoveries based on these methods is low. Furthermore, these methods fail to grasp the common molecular changes in cells transitioning from a normal state to the cancerous state. Choi et al. [22] introduced a model to find differential gene coexpression patterns related to cancer by combining independent datasets for different cancers. They used a model similar to the $t$-test, which only considered the mean and variance of two groups of samples. It is well known that traditional $t$-test has two disadvantages for microarray data analysis: first, it assumes that the datasets under analysis have a normal distribution, which is usually violated in microarray datasets; second, if the number of genes is large and the number of samples is small, some of the standard deviations will be extremely small, and therefore the test statistics will be very high, which may lead to a significant bias. Nonparametric statistical test methods, such as the K-S test, require fewer assumptions for the data and may be preferred, especially, when the number of samples is small.

In this paper, we propose a novel method to detect the differentially changed gene relations in cancer versus normal tissues. We collect 36 datasets across different microarray platforms and from various types of cancer. These 36 datasets contain both normal and tumor samples, which can subsequently yield two Pearson correlation coefficient vectors for every gene pair, one for normal samples and the other for tumor samples. We then perform a bootstrapping K-S test to identify some differentially changed gene relations. Finally we verify our results with three key pathways related to cancer and demonstrate that our method can find some meaningful alterations of gene relations.

## 2. Materials and Methods

*2.1. Microarray Datasets.* We collected 36 microarray datasets from NCBI (Gene Expression Omnibus GEO) [23]. As shown in Table 1, these microarray datasets contain both normal and tumor samples across 21 different types of cancer, and their platforms come from one of the three platforms: GPL570 (Affymetrix GeneChip Human Genome U133 Plus 2.0 Array), GPL96 (Affymetrix GeneChip Human Genome U133 Array Set HG-U133A), and GPL91 (Affymetrix GeneChip Human Genome U95 Version Set HG-U95A). We divided every dataset into two expression data matrices: one matrix includes all normal samples, and the other includes all tumor samples. To integrate multiple microarray datasets across different platforms, we mapped each probe in different platforms to a unique Entrez Gene ID or a unique UniGene symbol. For genes with more than one probe in one platform, we chose the probe with the highest mean expression value.

*2.2. Cancer-Associated Pathways and Extended Gene Networks.* We applied our method to analyze three cancer-associated pathways. These pathways are related to three common traits in most and perhaps all types of human cancer: self-sufficiency in growth signals, insensitivity to antigrowth signals, and evading programmed cell death (apoptosis) [24]. In fact, Hanahan and Weinberg have already identified some signaling pathways to demonstrate the capabilities cancer cells acquire during tumor development in [24]. We extended these signaling pathways to three relatively complete and larger cancer-associated pathways (antigrowth signaling, apoptosis, and growth signaling pathways) from the cell cycle pathway, the apoptosis pathway and the MAPK pathway in KEGG [25]. We used these three pathways (i.e., cell cycle, apoptosis, and MAPK pathways) as our seeds and the genes in these pathways as our seed genes. Next we constructed three gene networks corresponding to the three cancer-associated pathways from HPRD (Human Proteins Reference Database, http://www.hprd.org/) and TRANSFAC [26] based on seed genes and their interacting partners. We downloaded the protein-protein interaction (PPI) data released by HPRD on September 1, 2007. This PPI dataset contains 37107 human binary protein-protein interactions whose supporting experiments are indicated as in vivo, in vitro, or yeast two-hybrid. We also collected 1042 transcription factor-target gene relations on human species from TRANSFAC. So our gene networks included seed genes, protein interaction partners, and transcription factors (TFs) of seed genes or target genes for which seed genes served as their TFs.

*2.3. Detecting Differential Relations by Bootstrapping K-S Test.* We used the Kolmogorov-Smirnov test (K-S test) to determine whether the distributions of values in two datasets differed significantly. The two-sample K-S test is the most useful for comparing two samples because it is nonparametric and distribution-free [27]. The null hypothesis for this test is that two datasets are drawn from the same distribution. The alternative hypothesis is that they are drawn from different distributions.

For $n$ i.i.d samples $X_1, \ldots, X_n$ with some unknown distribution, we can define an empirical distribution function by

$$S_n(x) = \begin{cases} 0, & \text{if } x < X_{(1)}, \\ \dfrac{k}{n}, & \text{if } X_{(k)} \le x < X_{(k+1)} \quad \text{for } k = 1, 2, \ldots, n-1, \\ 1, & \text{if } x \ge X_{(n)}, \end{cases} \tag{1}$$

where $X_1, \ldots, X_n$ are ordered from the smallest to the largest value. The Kolmogorov-Smirnov statistic for a given function $S(x)$ is

$$D_n = \max_x |S_n(x) - S(x)|. \tag{2}$$

$D_n$ will converge to 0 if the sample comes from distribution $S(x)$ [27]. Moreover, the cumulative distribution function of

Table 1: List of 36 microarray datasets.

| Series ID in GEO | Cancer type | Numbers of normal samples | Numbers of tumor samples | Numbers of genes | Platform ID in GEO |
|---|---|---|---|---|---|
| GSE3744 | Breast cancer | 7 | 40 | 54681 | GPL570 |
| GSE5764 | Breast cancer | 20 | 10 | 54681 | GPL570 |
| GSE7904 | Breast cancer | 19 | 43 | 54681 | GPL570 |
| GSE3678 | Thyroid cancer | 7 | 7 | 54681 | GPL570 |
| GSE3467 | Thyroid cancer | 9 | 9 | 54681 | GPL570 |
| GSE8977 | Breast cancer | 15 | 7 | 54681 | GPL570 |
| GSE8671 | Colorectal cancer | 32 | 32 | 54681 | GPL570 |
| GSE4290 | Glioma | 23 | 157 | 54681 | GPL570 |
| GSE4183 | Colorectal cancer | 8 | 30 | 54681 | GPL570 |
| GSE4107 | Colorectal cancer | 10 | 12 | 54681 | GPL570 |
| GSE8514 | Aldosterone-producing adenoma | 5 | 10 | 54681 | GPL570 |
| GSE6791 | Cervical cancer | 8 | 20 | 54681 | GPL570 |
| GSE6791 | Head and neck cancer | 18 | 38 | 54681 | GPL570 |
| GSE6338 | Lymphoma | 20 | 40 | 54681 | GPL570 |
| GSE5563 | Vulvar intraepithelial neoplasia | 9 | 9 | 54681 | GPL570 |
| GSE6004 | Thyroid Cancer | 4 | 14 | 54681 | GPL570 |
| GSE2549 | Malignant pleural mesothelioma | 10 | 44 | 22283 | GPL96 |
| GSE781 | Kidney cancer | 9 | 8 | 22283 | GPL96 |
| GSE7670 | Lung cancer | 27 | 27 | 22283 | GPL96 |
| GSE6344 | Kidney cancer | 10 | 10 | 22283 | GPL96 |
| GSE1542 | Pancreatic ductal carcinoma | 25 | 24 | 22283 | GPL96 |
| GSE6883 | Breast cancer | 6 | 6 | 22283 | GPL96 |
| GSE2724 | Uterine fibroid | 11 | 7 | 22283 | GPL96 |
| GSE2503 | Skin cancer | 6 | 5 | 22283 | GPL96 |
| GSE3268 | Lung cancer | 5 | 5 | 22283 | GPL96 |
| GSE9476 | Acute myeloid leukemia | 38 | 26 | 22283 | GPL96 |
| GSE6008 | Ovarian tumor | 4 | 99 | 22283 | GPL96 |
| GSE6477 | Multiple myeloma | 12 | 150 | 22283 | GPL96 |
| GSE4115 | Lung Cancer | 90 | 97 | 22283 | GPL96 |
| GSE3167 | Bladder cancer | 14 | 46 | 22283 | GPL96 |
| GSE2514 | Pulmonary adenocarcinoma | 19 | 20 | 12651 | GPL91 |
| GSE6631 | Head and neck cancer | 22 | 22 | 12651 | GPL91 |
| GSE6604<br>GSE6605 | Prostate tumor | 18 | 25 | 12651 | GPL91 |
| GSE6606<br>GSE6608 | Prostate tumor | 63 | 65 | 12651 | GPL91 |
| GSE2379 | Head and neck cancer | 4 | 34 | 12651 | GPL91 |
| GSE1987 | Lung Cancer | 9 | 28 | 12651 | GPL91 |

Kolmogorov distribution is

$$K(x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{i=1}^{\infty}e^{-(2i-1)^2\,\pi^2/(8x^2)}. \tag{3}$$

It is easy to prove that $\sqrt{n}D_n = \sqrt{n}\max_x|S_n(x) - S(x)|$ will converge to the Kolmogorov distribution [27]. Therefore if $\sqrt{n}D_n > K_\alpha = \Pr(K \le K_\alpha) = 1 - \alpha$, the null hypothesis for the Kolmogorov-Smirnov test will be rejected at level $\alpha$.

For the case of determining whether the distributions of two data vectors differ significantly, the Kolmogorov-Smirnov statistic is

$$D_{n,m} = \max_x |S_n(x) - S_m(x)|, \tag{4}$$

and the null hypothesis will be rejected at level $\alpha$ if

$$\sqrt{\frac{nm}{n+m}}D_{n,m} > K_\alpha. \quad (5)$$

The $P$-value from the K-S test can measure the confidence of the comparison results against the null hypothesis. Obviously, the smaller the $P$-value, the more confident we are of rejecting the null hypothesis.

Assume that we have $n$ microarray datasets and a list of $m$ genes, we denote the expression data matrix for normal samples as

$$N^k = \begin{pmatrix} X_{11}^k & X_{12}^k & \cdots & X_{1p}^k \\ X_{21}^k & X_{22}^k & \cdots & X_{2p}^k \\ . & . & . & . \\ X_{m1}^k & X_{m2}^k & \cdots & X_{mp}^k \end{pmatrix} \quad k = 1, \ldots, n, \quad (6)$$

and the expression data matrix for tumor samples as

$$T^l = \begin{pmatrix} Y_{11}^l & Y_{12}^l & \cdots & Y_{1q}^l \\ Y_{21}^l & Y_{22}^l & \cdots & Y_{2q}^l \\ . & . & . & . \\ Y_{m1}^l & Y_{m2}^l & \cdots & Y_{mq}^l \end{pmatrix} \quad l = 1, \ldots, n, \quad (7)$$

where $p(k)$ is the number of normal samples in the $k$th dataset, and $q(l)$ is the number of tumor samples in the $l$th dataset.

For these two types of expression data matrices, each row represents one gene, and each column represents one sample. The correlation coefficient for gene $i$ and gene $j$ from the $k$th normal sample can be calculated by

$$\mathrm{NPC}_{ij}^k = \frac{\sum_{a=1}^{p}\left(X_{ia}^k - \overline{X}_i^k\right)\left(X_{ja}^k - \overline{X}_j^k\right)}{\sqrt{\sum_{a=1}^{p}\left(X_{ia}^k - \overline{X}_i^k\right)^2}\sqrt{\sum_{a=1}^{p}\left(X_{ja}^k - \overline{X}_j^k\right)^2}}, \quad (8)$$

where $\overline{X}_i^k$ is the average value of expression levels for gene $i$. The correlation coefficient for every gene pair from tumor samples can be calculated similarly.

We use the bootstrapping K-S test to detect some gene relations with different PC (Pearson coefficient) distributions. The bootstrapping method generates $N$ bootstrapping samples NPC and TPC by repeatedly sampling with replacement from the original $\mathrm{NPC}_{ij}$ and $\mathrm{TPC}_{ij}$ (e.g., Step 4), respectively. It can give us an empirical distribution of $P$-value $\theta$, with which, we can estimate the probability that the distribution of two PC vectors are different. In our computational experiment, for a gene pair, if its value of $\Pr(\theta < 0.05)$ was larger than 0.8, we considered it as a pair of genes with the correlation relation significantly different between normal and cancer cells.

Our method can be described as follows.

*Step 1.* Compute $n$ correlation coefficient Matrices $\mathrm{NPC}^1$–$\mathrm{NPC}^n$ from the normal samples in $n$ datasets for every gene pairs. For example, $\mathrm{NPC}^1$ is an $m \times m$ Matrix from normal samples in the first dataset, and $\mathrm{NPC}_{ij}^1$ represents the correlation coefficient between gene $i$, and gene $j$.

*Step 2.* Compute $n$ correlation coefficient Matrices $\mathrm{TPC}^1$–$\mathrm{TPC}^n$ from the tumor samples in the $n$ datasets for every gene pair.

*Step 3.* For every gene pair (gene $i$ and gene $j$), let

$$\begin{aligned} \mathrm{NPC}_{ij} &= \left[\mathrm{NPC}_{ij}^1 \;\; \mathrm{NPC}_{ij}^2 \;\; \mathrm{NPC}_{ij}^3 \;\; \cdots \;\; \mathrm{NPC}_{ij}^n\right], \\ \mathrm{TPC}_{ij} &= \left[\mathrm{TPC}_{ij}^1 \;\; \mathrm{TPC}_{ij}^2 \;\; \mathrm{TPC}_{ij}^3 \;\; \cdots \;\; \mathrm{TPC}_{ij}^n\right], \end{aligned} \quad (9)$$

*Step 4.* Perform the following ($N$ is the number of samples we will generate using bootstrapping).

*for* $k = 1$ to $N$

**Do** generate bootstrap samples NPC and TPC from $\mathrm{NPC}_{ij}$ and $\mathrm{TPC}_{ij}$, respectively.

$\theta_k$ = $P$-value of K-S test on NPC and TPC.

End-*for*

Output $\Pr(\theta < 0.05) = \sharp\ (\theta < 0.05)/N$.

## 3. Experimental Results

In this section, we applied the bootstrapping K-S test method to analyze three cancer related pathways.

*3.1. Antigrowth Signaling Pathway.* Antigrowth signals can control proliferation in normal samples. Cancer cells have the ability to evade these antiproliferation signals. In the antigrowth signaling pathway, transforming growth factor beta (TGF$\beta$) initiates this pathway by binding to two TGF$\beta$ receptors, Tgfbr1 and Tgfbr2. These two activated Tgf$\beta$ receptors can phosphorylate Smad2, Smad3, and Smad4 [28]. The SMAD family proteins then transduce antigrowth signals to the cell cycle inhibitors p21, p16, p27, and p15, which can inhibit the action of cyclin-CDK complex. The cyclin-CDK complex can phosphorylate RB and make RB dissociate from the E2F/RB complex to liberate E2F to activate the cell cycle procession from G1 to S phase (Figure 1(a)).

There are 19 genes in the antigrowth signaling pathway (Figure 1(a)). We found 689 unique genes related to these 19 genes from TRANSFAC and HPRD. Among these 708 genes, there were 4215 paired gene interactions, among which the correlation relations of 47 gene pairs were identified as significantly changed between normal and cancer cells. Among these 47 relations, we detected a cluster around SMAD family proteins which contained 15 relations with different distributions between normal samples and tumor samples (Figure 1(b)). Most of them came from large-scale protein-protein interaction experiments without the associated molecular function. For example, (Smad1–Arl4d), (RHOD–Smad2), and (WEE1–Smad3) in [29], (PAPOLA–Smad2), (SNRP70–Smad5), (GPNMB–Smad4), (PSMD11–Smad3), and (Smad9–MBD1) in [30], and (EWSR1–Smad4) in [31], all of them were detected based on large-scale protein-protein interaction experiments without annotation

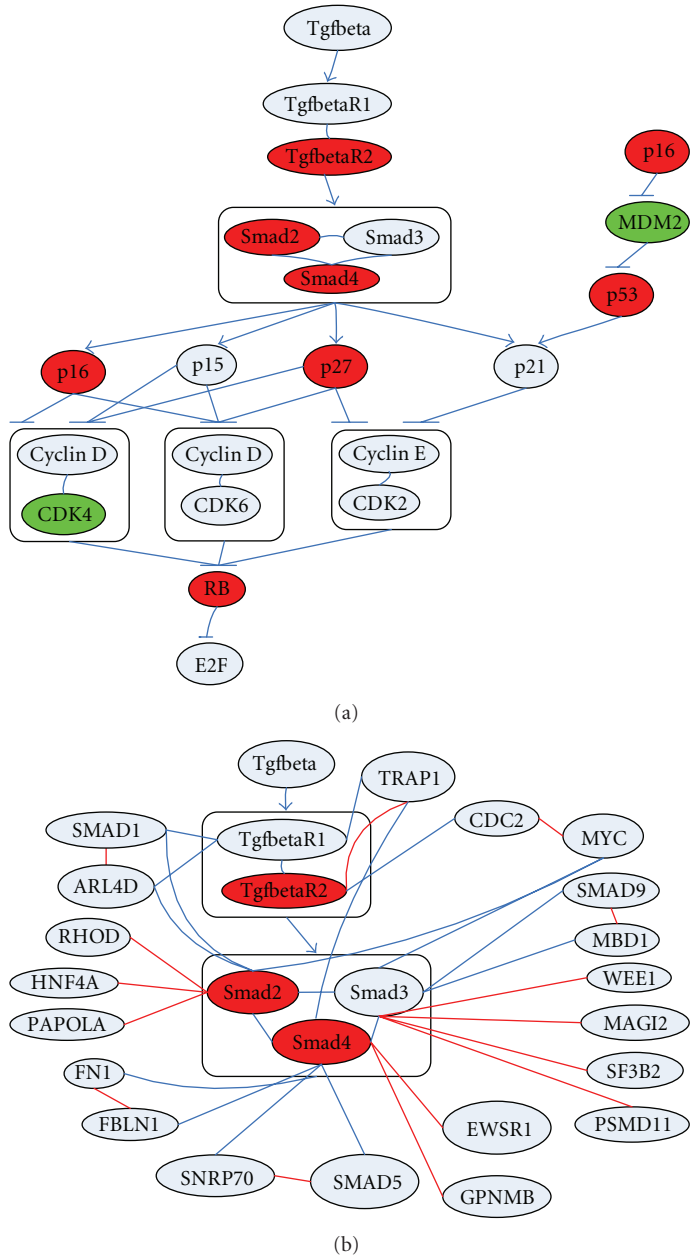


Figure 1: Antigrowth signaling pathway and cluster around SMAD proteins. (a) Antigrowth signaling pathway. Nodes and edges represent human proteins and protein-protein interactions, respectively. Edges with direction represent a regulatory relation. → means an activating relation and, ⊣ means an inhibitory relation. (b) Cluster around smads. Red edges represent differentially changed relations. Blue edges represent unchanged relations. Red nodes represent tumor suppressor genes, and green nodes represent oncogenes.

of molecular function. Our results indicate that although their associated functions and internal mechanisms are still unclear, these gene pairs are related to the TGF$\beta$-SMAD signaling pathway, and the relation between the two genes in each pair is significantly different in cancer and normal cells. Additionally, we identified some differentially changed relations with known molecular functions as follows:

(1) MAGI2 (a.k.a. ARIP1)–Smad3. MAGI2 (ARIP1) can interact with Smad3, and overexpression of ARIP1 can significantly suppress Smad3-induced transcriptional activity [32]. We validated this from the boxplot for MAGI2 (ARIP1)–Smad3 (Figure 2(a)). In normal samples, MAGI2 (ARIP1) and Smad3 showed a high positive correlation, while they had a high negative correlation in tumor samples.

(2) EWSR1–Smad4. Although the experiment type of the interaction between EWSR1 and Smad4 is yeast two-hybrid [31], mutations in EWSR1 are known to cause Ewing sarcoma and other members of the

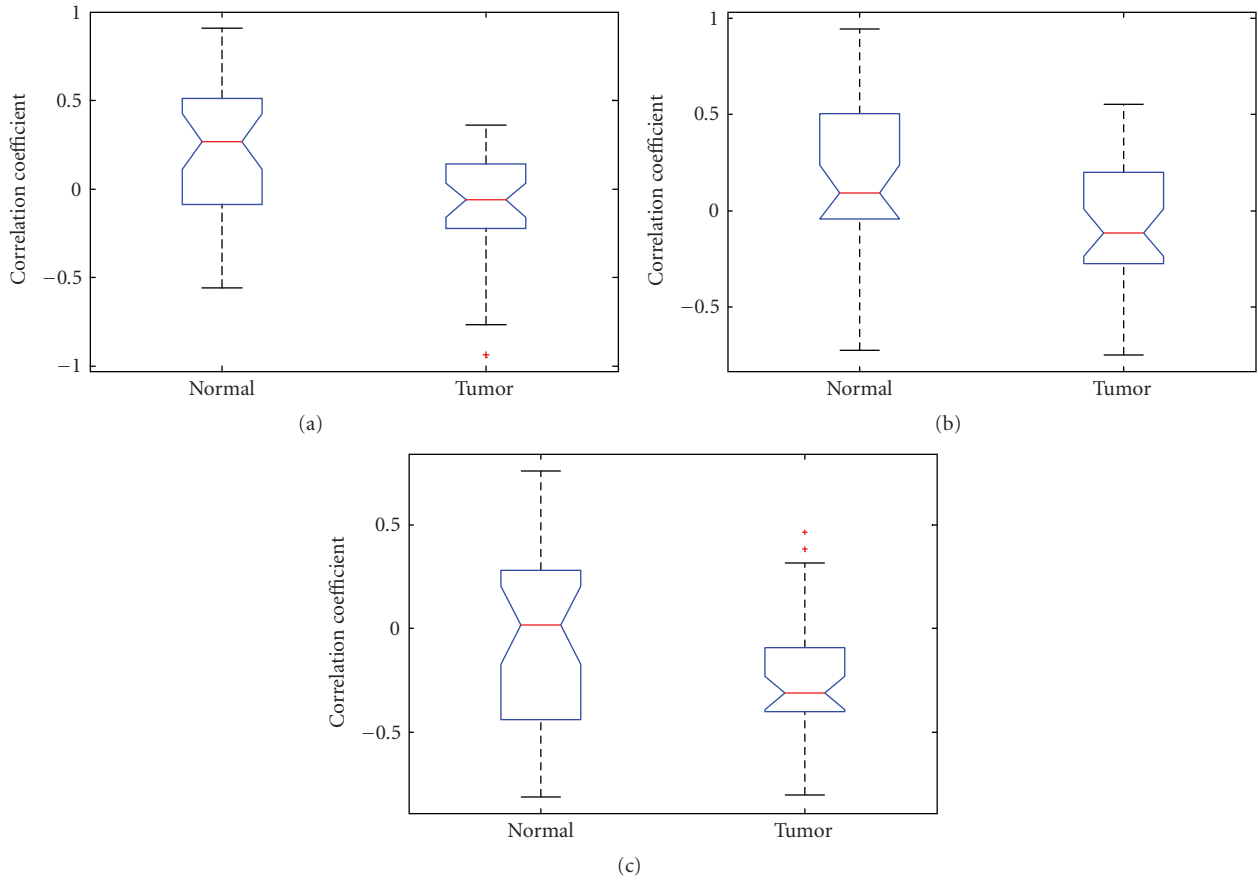


Figure 2: (a) Boxplot for MAGI2 (ARIP1)–Smad3. $\Pr(\theta < 0.05) = 0.986$. (b) Boxplot for EWSR1–Smad4. $\Pr(\theta < 0.05) = 0.954$. (c) Boxplot for TRAP1–TgfbetaR2. $\Pr(\theta < 0.05) = 0.944$.

Ewing family of tumors [33]. From the boxplot for EWSR1–Smad4, we found that the third quartile is the densest part of the whole distribution for both normal and tumor samples. The third quartile for normal samples showed a positive correlation whereas that for tumor samples showed a negative correlation (Figure 2(b)). Therefore, we suspect that EWSR1 can suppress the activity of Smad4 in tumor samples.

(3) TRAP1–Tgfbr2. TRAP1 has been shown to bind to TGF$\beta$ receptors and play a role in TGF$\beta$ signaling pathway. TRAP1 can interact with Smad4 and affect the SMAD-mediated signal transduction pathway. Mutant TRAP1 can prevent the formation of the Smad2–Smad4 complex to inhibit the TGF$\beta$ Signaling pathway [34]. In the boxplot for TRAP1–Tgfbr2 (Figure 2(c)), the densest quartile for tumor samples showed a high negative correlation.

*3.2. Apoptosis Pathway.* Cancer cells have the ability to evade programmed cell death or apoptosis. TNF$\alpha$, FASL, TRAIL, and other genes can initiate apoptosis by binding to their receptors such as TNFR1, FAS, and TRAIL-R. Many apoptosis signals induce mitochondrial changes. Mitochondria can help transduce the apoptosis signals by releasing cytochrome C (Cytc), a potent catalyst of apoptosis. There are two different Bcl-2 family members: proapoptotic members (Bid, BAD) and antiapoptotic members (Bcl-2, Bcl-xl), which activate and inhibit, respectively, the release of Cytc. Finally, two key caspases (Casp8 and Casp9) activate other downstream caspases that perform the cascading events of cell death (Figure 3(a)).

In our results, we detected 33 relations with different distributions in the apoptosis pathway, and some are supported by existing experimental evidence. Examples include (Figure 3(b)) the following:

(1) PUMA–Bcl-XL (BCL2L1). PUMA can interact with Bcl-XL and meanwhile PUMA can also neutralize and antagonize all the Bcl-2-like proteins [35]. From the boxplot for PUMA–Bcl-XL, we can find that Bcl-XL, and PUMA showed a higher negative correlation in normal samples than in tumor samples (Figure 4(a)).

(2) AKT1–BAD. Active forms of Akt can phosphorylate BAD in vivo and in vitro to prevent it from promoting cell death [36]. In the boxplot for AKT1–BAD, the first quartile, the densest quartile for normal samples, showed a higher positive correlation than the second

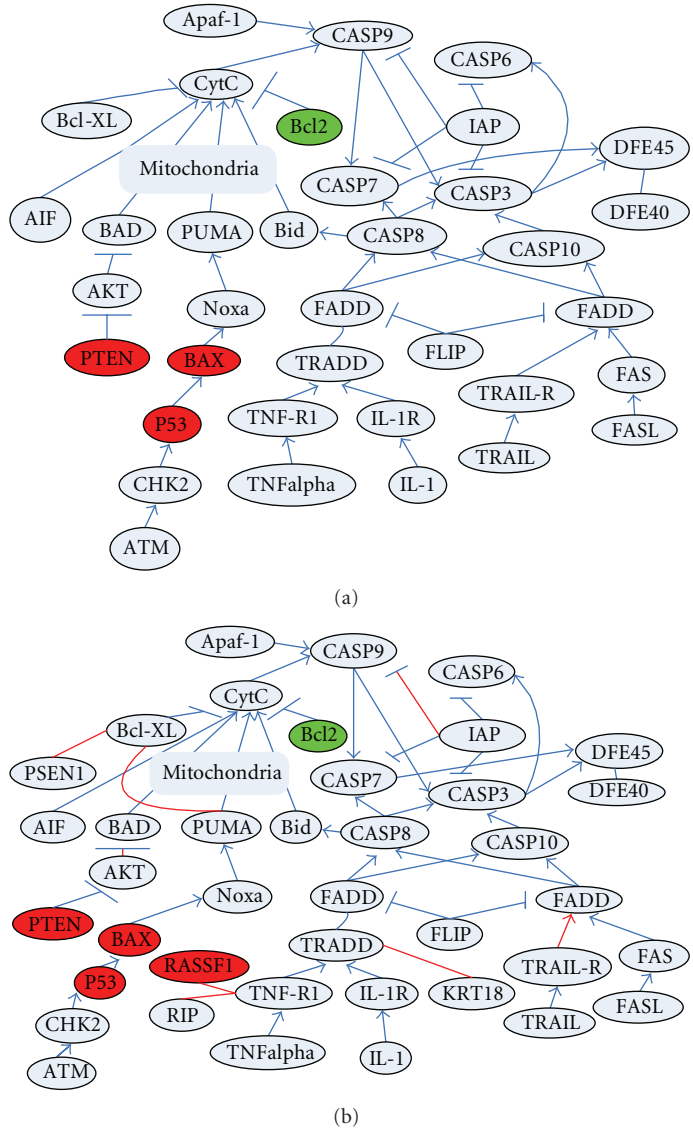


FIGURE 3: (a) Apoptosis pathway. (b) Differentially changed gene relations in apoptosis pathway. Red edges represent differentially changed relations. Blue edges represent unchanged relations. Red nodes represent tumor suppressor genes, and green nodes represent oncogenes.

quartile, the densest for tumor samples (Figure 4(b)). So we speculated that Akt can suppress BAD's activity in tumor samples.

(3) KRT18–TRADD. TRADD is a KRT18-interacting protein. KRT18 may inactivate TRADD to prevent interactions between TRADD and the activated TNFR1 and thus affect TNF$\alpha$-induced apoptosis [37]. In the boxplot for KRT18–TRADD, normal samples showed a higher positive correlation (Figure 4(c)).

(4) TNFR1–RIPK1 (RIP). The interaction between the death domain of TNF$\alpha$ receptor-1 (TNFR1) and TRADD can trigger distinct signaling pathways leading to apoptosis. TRADD also interacts strongly with another death domain protein; RIP and RIP plays an important role in the TNF signaling cascades leading to apoptosis [38]. In the boxplot for TNFR1–RIPK1, TNFR1 and RIPK1 exhibited high positive correlation in normal samples (Figure 4(d)).

(5) TNFR1–RASSF1. RASSF1A is a tumor suppressor gene. Apoptosis initiation by TNF$\alpha$ or TRAIL recruits RASSF1A and MAP-1 to form complexes. RASSF1A and MAP-1 are the key links between death receptors and the apoptotic machinery [39]. This was verified by the Boxplot for TNFR1–RASSF1. In most normal samples, these genes showed a high positive correlation. In most tumor samples, they showed a zero or negative correlation (Figure 4(e)).

(6) IAP–CASP9. Inhibitor of apoptosis (IAP) suppresses the activities of caspases and inhibits different apoptotic pathways [40]. IAP and CASP9 showed a high negative correlation in tumor samples (Figure 4(f)).

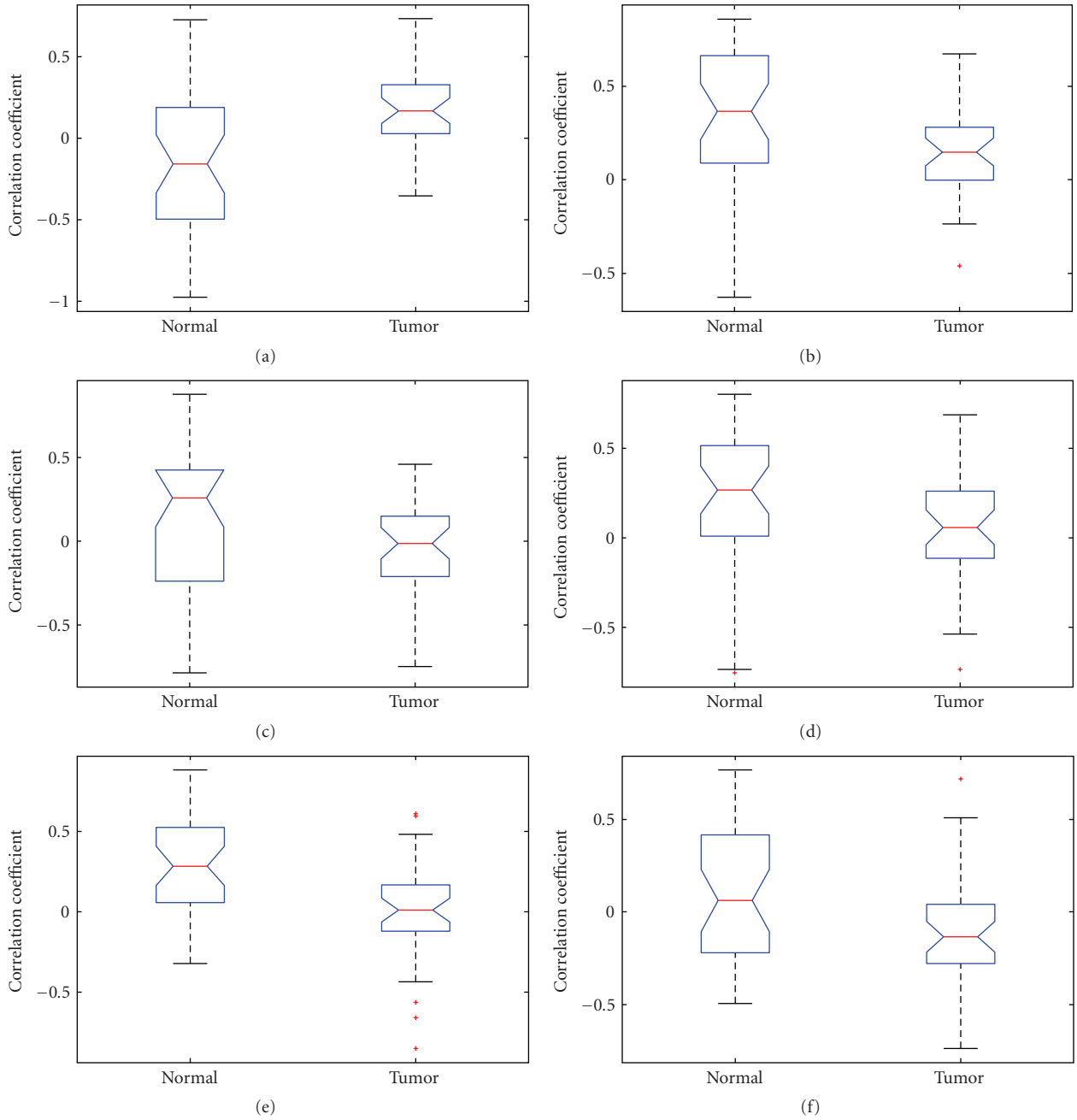


Figure 4: (a) Boxplot for PUMA–Bcl-XL(BCL2L1). $\Pr(\theta < 0.05) = 0.998$. (b) Boxplot for AKT1–BAD. $\Pr(\theta < 0.05) = 0.859$. (c) Boxplot for KRT18–TRADD. $\Pr(\theta < 0.05) = 0.991$.(d) Boxplot for TNFR1–RIPK1(RIP). $\Pr(\theta < 0.05) = 0.831$. (e) Boxplot for TNFR1–RASSF1. $\Pr(\theta < 0.05) = 0.946$. (f) Boxplot for IAP–CASP9. $\Pr(\theta < 0.05) = 0.826$.

Among the eight differential gene relations in Figure 3(b), three of them were in the seed pathway: TRAIL-R → FADD, IAP → CASP9, and AKT → BAD, which demonstrates the effectiveness of the proposed method.

*3.3. Growth Signaling Pathway.* Cancer cells have the ability to produce their own growth promoting signals. EGF, TGF$\alpha$, and PDGF are activated and then bind to their receptors to transduce the growth signals. The activated growth factor receptors can in turn activate the SOS-Ras_Raf_Mapk cascade. In the growth signal pathway (Figure 5), Ras, JUN, and Fos are oncogenes.

We could find 68 relations with different distributions in the growth signal pathway, and we discuss three relations as follows:

(1) RASSF2–KRAS. Although different forms of Ras are frequently thought of as oncogenes, they also have the ability to produce antigrowth effects such as cell

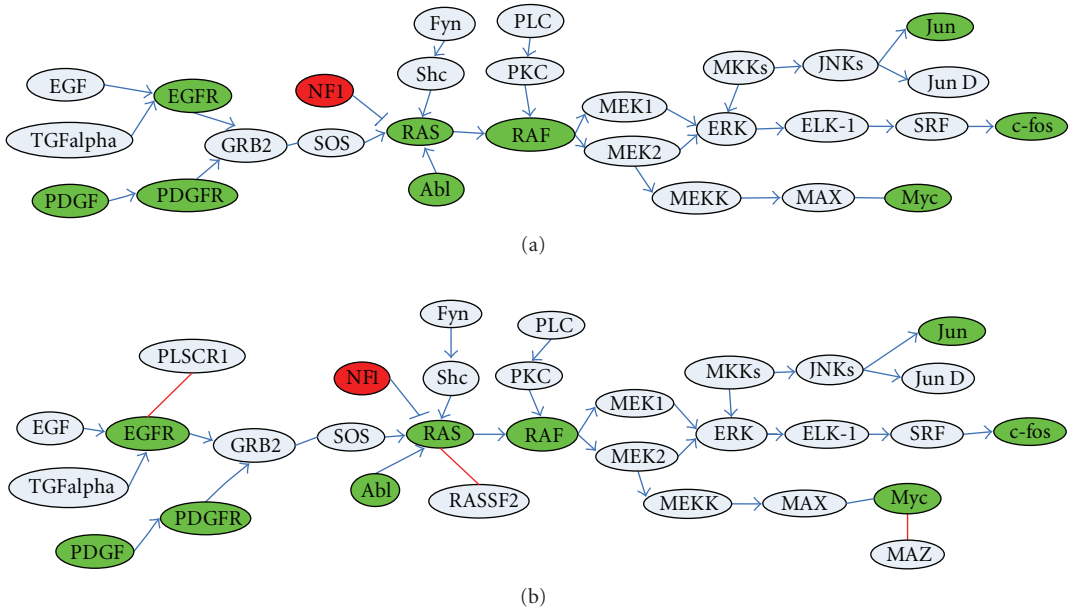


FIGURE 5: (a) Growth signal pathway. (b) Differentially changed relations in growth signal pathway. Red edges represent differentially changed relations. Blue edges represent unchanged relations. Red nodes represent tumor suppressor genes, and green nodes represent oncogenes.

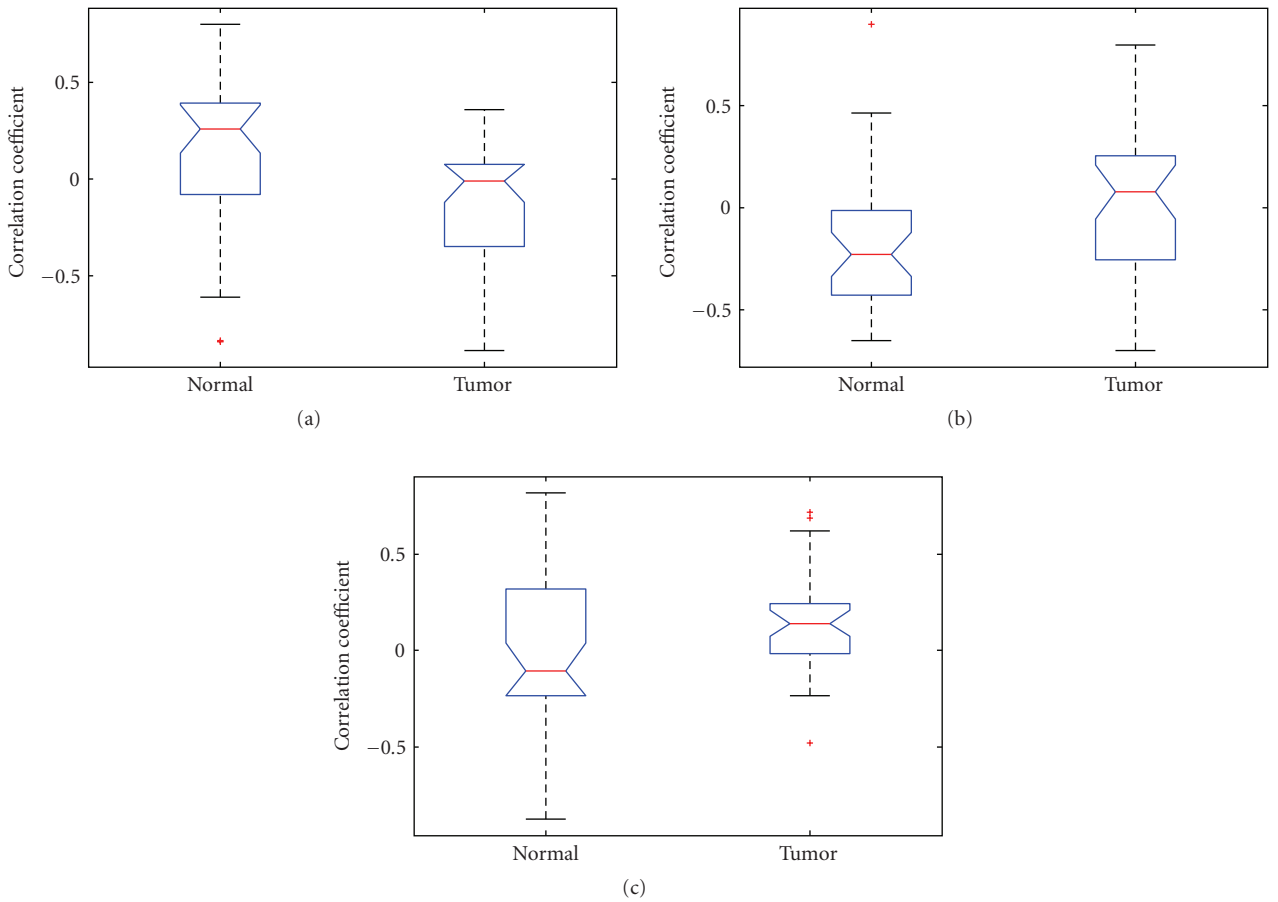


FIGURE 6: (a) Boxplot for RASSF2–KRAS. $\Pr(\theta < 0.05) = 0.983$. (b) Boxplot for MAZ–MYC. $\Pr(\theta < 0.05) = 0.833$. (c) Boxplot for PLSCR1–EGFR. $\Pr(\theta < 0.05) = 0.963$.

cycle arrest, differentiation, and apoptosis. RASSF2 can bind directly to K-Ras. Moreover, RASSF2 can inhibit the growth of tumor cells, and the activated K-Ras can enhance this ability [41]. This might be why RASSF2 and RAS showed a high positive correlation in normal samples in the boxplot (Figure 6(a)).

(2) MAZ–MYC. The MAZ family can increase the oncogene MYC's transcriptional activity [42]. As expected, MAZ and MYC demonstrated a higher positive correlation in tumor samples (Figure 6(b)).

(3) PLSCR1–EGFR. Activated epidermal growth factor receptors (EGFRs) can both physically and functionally interact with PLSCR1. In turn, PLSCR1 can interact with Shc and thus accelerate the activation of Src kinase through the EGF receptor, while Src can initiate some activating pathway for the oncogene JUN [43]. In the boxplot for PLSCR1–EGFR, the densest quartile for normal samples showed a low negative correlation, whereas the densest quartile for tumor samples showed a low positive correlation (Figure 6(c)).

## 4. Conclusion and Discussion

After several decades of cancer research, some details of the underlying mechanisms of cancer at the gene level are still unclear. In this paper, we propose an integrative method based on the bootstrapping K-S test to evaluate a large number of microarray datasets generated from 21 different types of cancer in order to identify gene pairs that have different relationships in normal versus cancer tissues. The significant alteration of gene relations can greatly extend our understanding of the molecular mechanisms of human cancer. In our method, we obviate the disadvantage of the traditional $t$-test, which only considers the mean and variance of samples and fails in the analysis of microarray data with small numbers of samples. Instead of the $t$-test, we propose the use of the bootstrapping K-S test method to detect gene pairs with different distributions of Pearson correlation coefficient values in normal and tumor samples. The experimental results demonstrated that our method could find meaningful alterations in gene relations and opened a potential door for further cancer research.

## Acknowledgment

This work was supported by NSF award IIS-0644366.

## References

[1] H. Han, D. J. Bearss, L. W. Browne, R. Calaluce, R. B. Nagle, and D. D. Von Hoff, "Identification of differentially expressed genes in pancreatic cancer cells using cDNA microarray," *Cancer Research*, vol. 62, no. 10, pp. 2890–2896, 2002.

[2] X.-W. Chen, "Margin-based wrapper methods for gene identification using microarray," *Neurocomputing*, vol. 69, no. 16–18, pp. 2236–2243, 2006.

[3] A. A. Margolin, I. Nemenman, K. Basso, et al., "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, supplement 1, article S7, pp. 1–15, 2006.

[4] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[5] X.-W. Chen, G. Anantha, and X. Wang, "An effective structure learning method for constructing gene networks," *Bioinformatics*, vol. 22, no. 11, pp. 1367–1374, 2006.

[6] H. Xiong and X.-W. Chen, "Kernel-based distance metric learning for microarray data classification," *BMC Bioinformatics*, vol. 7, article 299, pp. 1–11, 2006.

[7] T. Golub, D. Slonim, P. Tamayo, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[8] H. Xiong, Y. Zhang, and X.-W. Chen, "Data-dependent kernel machines for microarray data classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 4, pp. 583–595, 2007.

[9] T.-M. Chu, B. Weir, and R. Wolfinger, "A systematic statistical linear modeling approach to oligonucleotide array experiments," *Mathematical Biosciences*, vol. 176, no. 1, pp. 35–51, 2002.

[10] W.-P. Hsieh, T.-M. Chu, R. D. Wolfinger, and G. Gibson, "Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles," *Genetics*, vol. 165, no. 2, pp. 747–757, 2003.

[11] M. Neuhäuser and R. Senske, "The Baumgartner-Weiβ-Schindler test for the detection of differentially expressed genes in replicated microarrays experiments," *Bioinformatics*, vol. 20, no. 18, pp. 3553–3564, 2004.

[12] O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman, "Nonparametric methods for identifying differentially expressed genes in microarray data," *Bioinformatics*, vol. 18, no. 11, pp. 1454–1461, 2002.

[13] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[14] H. Yang and G. Churchill, "Estimating $p$-values in small microarray experiments," *Bioinformatics*, vol. 23, no. 1, pp. 38–43, 2007.

[15] Y. Zhao and W. Pan, "Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 19, no. 9, pp. 1046–1054, 2003.

[16] S. Draghici, P. Khatri, A. L. Tarca, et al., "A systems biology approach for pathway level analysis," *Genome Research*, vol. 17, no. 10, pp. 1537–1545, 2007.

[17] T. Manoli, N. Gretz, H.-J. Gröne, M. Kenzelmann, R. Eils, and B. Brors, "Group testing for pathway analysis improves comparability of different microarray datasets," *Bioinformatics*, vol. 22, no. 20, pp. 2500–2506, 2006.

[18] K.-C. Li, "Genome-wide coexpression dynamics: theory and application," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 26, pp. 16875–16880, 2002.

[19] Y. Lai, B. Wu, L. Chen, and H. Zhao, "A statistical method for identifying differential gene-gene co-expression patterns," *Bioinformatics*, vol. 20, no. 17, pp. 3146–3155, 2004.

[20] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: is there a unique set?" *Bioinformatics*, vol. 21, no. 2, pp. 171–178, 2005.

[21] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nature Reviews Genetics*, vol. 7, no. 1, pp. 55–65, 2006.

[22] J. K. Choi, U. Yu, O. J. Yoo, and S. Kim, "Differential coexpression analysis using microarray data and its application to human cancer," *Bioinformatics*, vol. 21, no. 24, pp. 4348–4355, 2005.

[23] T. Barrett, T. O. Suzek, D. B. Troup, et al., "NCBI GEO: mining millions of expression profiles—database and tools," *Nucleic Acids Research*, vol. 33, database issue, pp. D562–D566, 2005.

[24] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, 2000.

[25] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.

[26] V. Matys, E. Fricke, R. Geffers, et al., "TRANSFAC®: transcriptional regulation, from patterns to profiles," *Nucleic Acids Research*, vol. 31, no. 1, pp. 374–378, 2003.

[27] W. J. Conover, *Practical Nonparametric Statistics*, John Wiley & Sons, New York, NY, USA, 3rd edition, 1999.

[28] G. C. Blobe, X. Liu, S. J. Fang, T. How, and H. F. Lodish, "A novel mechanism for regulating transforming growth factor $\beta$ (TGF-$\beta$) signaling: functional modulation of type III TGF-$\beta$ receptor expression through interaction with the PDZ domain protein, GIPC," *The Journal of Biological Chemistry*, vol. 276, no. 43, pp. 39608–39617, 2001.

[29] M. Barrios-Rodiles, K. R. Brown, B. Ozdamar, et al., "High-throughput mapping of a dynamic signaling network in mammalian cells," *Science*, vol. 307, no. 5715, pp. 1621–1625, 2005.

[30] F. Colland, X. Jacq, V. Trouplin, et al., "Functional proteomics mapping of a human signaling pathway," *Genome Research*, vol. 14, no. 7, pp. 1324–1332, 2004.

[31] J.-F. Rual, K. Venkatesan, T. Hao, et al., "Towards a proteome-scale map of the human protein-protein interaction network," *Nature*, vol. 437, no. 7062, pp. 1173–1178, 2005.

[32] H. Shoji, K. Tsuchida, H. Kishi, et al., "Identification and characterization of a PDZ protein that interacts with activin type II receptors," *The Journal of Biological Chemistry*, vol. 275, no. 8, pp. 5485–5492, 2000.

[33] J. Ban, C. Siligan, M. Kreppel, D. Aryee, and H. Kovar, "EWS-FLI1 in Ewing's sarcoma: real targets and collateral damage," *Advances in Experimental Medicine and Biology*, vol. 587, pp. 41–52, 2006.

[34] J. U. Wurthner, D. B. Frank, A. Felici, et al., "Transforming growth factor-$\beta$ receptor-associated protein 1 is a Smad4 chaperone," *The Journal of Biological Chemistry*, vol. 276, no. 22, pp. 19495–19502, 2001.

[35] L. Chen, S. N. Willis, A. Wei, et al., "Differential targeting of prosurvival Bcl-2 proteins by their BH3-only ligands allows complementary apoptotic function," *Molecular Cell*, vol. 17, no. 3, pp. 393–403, 2005.

[36] L. del Peso, M. González-García, C. Page, R. Herrera, and G. Nuñez, "Interleukin-3-induced phosphorylation of BAD through the protein kinase Akt," *Science*, vol. 278, no. 5338, pp. 687–689, 1997.

[37] H. Inada, I. Izawa, M. Nishizawa, et al., "Keratin attenuates tumor necrosis factor-induced cytotoxicity through association with TRADD," *The Journal of Cell Biology*, vol. 155, no. 4, pp. 415–426, 2001.

[38] H. Hsu, J. Huang, H.-B. Shu, V. Baichwal, and D. V. Goeddel, "TNF-dependent recruitment of the protein kinase RIP to the TNF receptor-1 signaling complex," *Immunity*, vol. 4, no. 4, pp. 387–396, 1996.

[39] S. Baksh, S. Tommasi, S. Fenton, et al., "The tumor suppressor RASSF1A and MAP-1 link death receptor signaling to bax conformational change and cell death," *Molecular Cell*, vol. 18, no. 6, pp. 637–650, 2005.

[40] Q. L. Deveraux, N. Roy, H. R. Stennicke, et al., "IAPs block apoptotic events induced by caspase-8 and cytochrome c by direct inhibition of distinct caspases," *The EMBO Journal*, vol. 17, no. 8, pp. 2215–2223, 1998.

[41] M. D. Vos, C. A. Ellis, C. Elam, A. S. Ülkü, B. J. Taylor, and G. J. Clark, "RASSF2 is a novel K-Ras-specific effector and potential tumor suppressor," *The Journal of Biological Chemistry*, vol. 278, no. 30, pp. 28045–28051, 2003.

[42] H. Tsutsui, O. Sakatsume, K. Itakura, and K. K. Yokoyama, "Members of the MAZ family: a novel cDNA clone for MAZ from human pancreatic islet cells," *Biochemical and Biophysical Research Communications*, vol. 226, no. 3, pp. 801–809, 1996.

[43] M. Nanjundan, J. Sun, J. Zhao, Q. Zhou, P. J. Sims, and T. Wiedmer, "Plasma membrane phospholipid scramblase 1 promotes EGF-dependent activation of c-Src through the epidermal growth factor receptor," *The Journal of Biological Chemistry*, vol. 278, no. 39, pp. 37413–37418, 2003.